CLAIMS

What is claimed is:

1. A method of collaborative focused crawling of documents related to multiple focus topics on a network, the method comprising:

selectively prioritizing the documents to crawl based on a set of rules;

fetching prioritized documents from the network;

for each fetched document, determining whether the fetched document is relevant to any of the multiple focus topics;

crawling the fetched document that matches any of the multiple focus topics; and

further crawling out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest.

2. The method of claim 1, further comprising seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents.

3. The method of claim 2, further comprising crawling the seed uniform resource locator strings.

4. The method of claim 3, further comprising writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings.

5. The method of claim 4, further comprising a foreman function for reading a plurality of contents of the resulting uniform resource locator strings.

6. The method of claim 5, further comprising the foreman function passing the contents of the resulting uniform resource locator strings to a miner.

7. The method of claim 6, further comprising the miner instructing a fetcher to crawl a plurality of out-links on a document of the resulting resource locator string when the contents of the resulting resource locator string match a focus topic of the miner.

8. The method of claim 6, further comprising the miner ignoring resulting resource locator string when the contents of the resulting resource locator string do not match the focus of the miner.

9. The method of claim 6, further comprising the miner managing a plurality of focus topics.

10. The method of claim 9, further comprising the miner allowing a crawling of the resulting resource locator string when the resulting resource locator string matches a plurality of web space rules.

11. The method of claim 9, wherein the web space rules comprise domain rules, IP address rules, and prefix rules.

12. The method of claim 9, further comprising the miner disallowing the crawling of the resulting resource locator string when the content of the resulting resource locator string matches a focus topic of the miner.

13. The method of claim 6, wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in a blacklist, so that the uniform resource locator strings will not be crawled

again.

14. A computer program product having instruction codes for implementing a collaborative focused crawling of documents related to multiple focus topics on a network, the computer program product comprising:

a first set of instruction codes for selectively prioritizing the documents to crawl based on a set of rules;

a second set of instruction codes for fetching prioritized documents from the network;

for each fetched document, a third set of instruction codes determines whether the fetched document is relevant to any of the multiple focus topics;

a fourth set of instruction codes for crawling the fetched document that matches any of the multiple focus topics; and

wherein the fourth set of instruction codes further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest.

15. The computer program product of claim 14, further comprising a fifth set of instruction codes for seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents.

16. The computer program product of claim 15, wherein the fourth set of instruction codes further crawls the seed uniform resource locator strings.

17. The computer program product of claim 16, further comprising a sixth set of instruction codes for writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings.

18. A system for implementing a collaborative focused crawling of documents related to multiple focus topics on a network, the system comprising:

an evaluator that selectively prioritizes the documents to crawl based on a set of rules;

a fetcher that fetches prioritized documents from the network;

for each fetched document, a focus engine determines whether the fetched document is relevant to any of the multiple focus topics;

a crawler for crawling the fetched document that matches any of the multiple focus topics; and

wherein the crawler further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest.

19. The system of claim 14, further comprising a plurality of seed uniform resource locator strings that are used to initiate the collaborative focused crawling of the documents.

20. The system of claim 15, wherein the crawler further crawls the seed uniform resource locator strings.